Poster: Biological Databases – BD9

PPStruct: a database of plant protein stuctures and annotations

Potenza E(1), Coller E(2), Hirsh L(1), Di Domenico T(1), Cestaro A(2), Tosatto SCE(1)

(1) Department of Biomedical Sciences, Universita' di Padova, Padova (2) Computational Biology Department, Fondazione Edmund Mach, Trento

Motivation: During the last ten years, the development of high-throughput sequencing, has generated a huge amount of genome sequences. Giving biological meaning to this data depends entirely on the capacity to develop instruments for its interpretation and organization. Moreover, once the protein sequences have been identified, functional annotation requires dedicated usage of an enormous amount of bioinformatics resources and specialized databases. Sequence annotation is often inaccurate and reliable predictions can only be obtained by using structure based functional annotation methods. These methods require the three-dimensional structure of the identified proteins. The experimental solution of protein structures is very time consuming and cannot be applied to all proteins in a genome, but has to be replaced with computational homology models. It is currently estimated that well over half of the known protein sequences can be predicted in this way. Plant genomics, despite its importance, started later than animal genomics. Currently the are less than ten plant genomes available in a genome browsers and few more at the “draft genome” level. In light of this limited amount of available data, any consideration regarding peculiar plant characteristics has to be considered temporary and seen with caution. Plant genomes were so far mostly annotated by hand, with an enormous expenditure of financial and human resources. Genome annotation for plants has to transit from prevalently manual towards fully automated annotation, with possible manual supervision, and is in serious need for the creation of new tools to permit this transition.
Methods: PPStruct database and website was designed with a multi-tier architecture, using separate modules for data management, data processing and presentation functions. To simplify development and maintenance, all tiers handle the common JSON (JavaScript Object Notation) format, thereby eliminating the need for data conversion. The MongoDB database engine is used for data storage and Node.js as middleware between data and presentation. PPStruct exposes its resources through RESTful web services, by using the Restify library for Node.js. The Angular.js framework and Bootstrap library were selected to provide the overall look-and-feel. Additional information is added to entries by querying the PDB and UNIPROT web services. Currently the genomes available at the database were annotated for the following features: - Domain assignment: InterPro tools set (Hunter et al., 2009) - Secondary structure: fastSS (Walsh et al., unpublished) - Disordered regions: MobiDB (Di Domenico et al., Bioinformatics 2012) - Homology modelling: HOMER (URL: http://protein.bio.unipd.it/homer/)
Results: Here we present PPStruct, a pipeline and a database dedicated to plant functional annotation. Our effort takes into account several specific aspects exploiting plant differences. The protein structure level is brought into play with the aim to better explain the effects of phenotypic differences at the molecular level. Reliable models are built for each gene transcript identified and the models will be used to better define the function of each protein. PPStruct website is currently under development but will be available soon from URL: http://ppstruct.bio.unipd.it/
Contact email: emilio.potenza@bio.unipd.it